DOCUMENT RESUME

ED 191 905 TM 800 529

AUTHOR Frisbie, David A.
TITLE The Relative Difficulty Ratio--A Test and Item Index.
PUB DATE Apr 80
NOTE 13p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; *Item Analysis; *Measurement Techniques; Scores; *Test Items
IDENTIFIERS *Relative Difficulty Ratio

ABSTRACT

The development of a new technique, the Relative Difficulty Ratio (RDR), is described, as well as how it can be used to determine the difficulty level of a test so that meaningful inter-test difficulty comparisons can be made. Assumptions made in computing RDR include: 1) each item must be scored dichotomously with only one answer choice keyed as correct; and 2) all wrong responses are a result of guessing. RDR has an upper limit of 1.00 for both a test and an item when all examinees have a perfect test or item score. The lower limit is reached when all examinees miss all items. The RDR not only facilitates comparisons between tests (i.e., multiple choice and true-false), but also can be useful in judging the adequacy of the difficulty level of a single test. As RDR decreases from zero, the interpretation is that the test is more difficult than is desirable. The opposite interpretation should be made when the RDR approaches its upper bound from zero. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED191905

THE RELATIVE DIFFICULTY RATIO - A TEST AND ITEM INDEX

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

David A. Frisbie

Measurement and Research Division
Office of Instructional Resources
University of Illinois at Urbana-Champaign

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
D. Frisbie

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the Annual Meeting of the
National Council on Measurement in Education

Boston

April 1980

TM 800529

The Relative Difficulty Ratio--A Test and Item Index

How can the difficulty levels of two content-parallel tests be compared meaningfully when the tests are not necessarily equivalent in the psychometric sense? This question arises in several applied measurement contexts and is primarily of concern to those who study the effect of test item form on various test characteristics. For example, comparable groups are administered either a 125-item true-false test or a 70-item four-choice multiple choice test. How can the difficulty levels of these tests be compared meaningfully? Studies by Hughes and Trimble (1965), Frisbie (1974), Dryden and Frisbie (1975), and Ebel (1978) are examples of situations in which this problem has occurred. In each case the authors were unable to make difficulty comparisons or such comparisons were made inappropriately.

Raw score means are not directly comparable between tests which have different chance scores. Whether chance scores vary because of test length differences or because of differences in item forms, mean raw score comparisons are not relevant. Of course, if objective tests are score with a correction for guessing formula, the means of the formula scores have been adjusted for chance success and are useful for judging relative difficulty. A recent study by Huck (1978) illustrates the problem encountered when means from tests having different chance scores are compared. The purpose of Huck's study was to show how examinee knowledge of item difficulties would affect their test score. However, the difficulty levels represented by the test means in his Table 2 are misleading because the chance scores for the multiple choice, true-false, and completion items differ. Assuming that formula scoring was not done, all ANOVA computations should have been done on means adjusted for differences in chance

success on subtests of varying item format. In addition, the item difficulty information given to subjects was misleading in that the p-values for the different item types were not adjusted for differences in chance success. For example, it is not clear that subjects would have interpreted a p = .80 differently for a true-false and multiple choice item.

The purpose of this paper is to show how a new technique, the Relative Difficulty Ratio (RDR) can be used to describe the difficulty level of a test so that meaningful inter-test difficulty comparisons can be made. The development of the RDR will be described and its application will be illustrated at both the test and item level.

Relative Difficulty Ratio

Derivation

The ideal mean of a test is the point midway between the chance score and a perfect score (Ebel, 1974). Because its value is dependent on both chance success at the item level and test length, the ideal mean represents a logical standard for assessing the difficulty level of a given test. The RDR is merely the ratio of the actual test mean to the ideal mean. An efficient formula for calculating the RDR can be derived simply by using $\bar{X}$ to represent the actual test mean and $\bar{X}_I$ to represent the ideal mean.

$$RDR_{\bar{X}} = \bar{X}/\bar{X}_I \qquad (1)$$

Equation 1 can be rewritten to simplify computations by representing the ideal mean as

$$\bar{X}_I = (K + K/n)/2 \qquad (2)$$

where K is the number of test items and n is the number of choices per item.

By simplifying equation 2 to

$$\bar{X}_I = K(n + 1)/2n \tag{3}$$

the RDR can be rewritten as

$$RDR_{\bar{X}} = 2n\bar{X}/K(n + 1) \tag{4}$$

A somewhat undesirable property of $RDR_{\bar{X}}$ represented by equation 4 is that the upper bound varies with n while both the ideal value (1.00) and lower bound (0.00) are constant for all values of n. Since in most practical situations mean scores occur in the upper portion of their potential range, $RDR_{\bar{X}}$ would be more easily interpreted if its upper bound was 1.00 for all values of n. Such a scale conversion can be managed easily by representing $RDR_{\bar{X}}$ as in equation 5 rather than equation 1.

$$RDR_{\bar{X}} = (\bar{X} - \bar{X}_I)/(K - \bar{X}_I) \tag{5}$$

By using equation 3 and simplifying, equation 6 can be used to facilitate computations.

$$RDR_{\bar{X}} = [2n\bar{X} - K(n + 1)]/K(n - 1) \tag{6}$$

In a similar fashion, the RDR of an item can be derived which is parallel to equations 5 and 6. The mean of item i on a test, p_i, is the traditional difficulty index or proportion passing item i. The chance score, p_c, is 1/n or the probability of choosing the correct answer by blind guessing. The ideal item mean can be written as

$$P_I = (1 + 1/n)/2 \tag{7}$$

since the perfect score mean is unity. The RDR for an item, RDR_i, can be written as

$$RDR_i = (p_i - p_I)/(1 - p_I) \tag{8}$$

Equation 8 can be rewritten to parallel equation 6 by substituting for p_I in equation 8

$$RDR_i = [n(2p_i - 1) - 1]/(n - 1) \tag{9}$$

Characteristics

Several reasonable assumptions must be made in using equations 6 and 9 to compute relative test or item difficulty. All items must be scored dichotomously and only one answer choice can be keyed as correct per item. Obviously formula scores should not be used with equation 6. The typical assumptions made when using the correction for guessing must also be made here. That is, all wrong responses are a result of guessing and all answer choices are equally attractive to the guesser. To the extent that these latter two assumptions are violated in a specific situation, the effect would be an underestimate of both the mean chance score and the ideal mean and an overestimate of $RDR_{\bar{X}}$.

A number of measurement specialists would say that the correction for guessing assumptions are always violated and, therefore, it is not reasonable to make such assumptions. Certainly this issue clouds the interpretation of a single RDR index, but it becomes a non-issue when RDRs from content-parallel tests or items are compared. There is no reason to expect the effect of violating the assumptions to influence one type of item more than another, unless the tests being compared are not content-parallel. For example, a 40-item five-choice multiple choice test containing a significant number of implausible distractors might be revised to three-choice or true-false form. There is no point in comparing the difficulty levels of the original and revised versions because the original is so deficient.

RDR has an upper limit of 1.00 for both a test and an item when all examinees have a perfect test or item score. The lower limit of $(n+1)/(1-n)$ is reached when all examinees miss all items. The ideal value of RDR for norm-referenced tests and items is zero; a value of -1.00 is obtained with the chance test or item mean. The effective range of RDR varies directly with n,

and n can take on any rational value greater than zero. For example, the value of n to be used in equation 6 when a test is comprised of 10 true-false, 40 four-choice, and 20 five-choice items is 3.99 or 4. This value was computed by summing the products 2(10/70), 4(40/70) and 5 (20/70) for this 70-item test.

The RDR not only facilitates comparisons between tests but also can be useful in judging the adequacy of the difficulty level of a single test. When RDR = 0.00, the test is of optimum difficulty in a norm-referenced sense; the conditions for high item discrimination and test reliability are most favorable. As the ratio decreases from zero, the interpretation is that the test is more difficult than is desirable. The opposite interpretation should be made when the RDR approaches its upper bound from zero.

Applications

Test Difficulty

Suppose the mean of a 50-item four-choice multiple choice test was 38.4. The $RDR_{\bar{X}}$ for this test, 0.29, suggests the (norm-referenced) test was easier than it should have been, ideally. Though this figure may be somewhat difficult to interpret in an absolute sense, relative to the upper bound of 1.00 for this test, most test specialists would recommend that a number of the easiest items be revised. For example, an item from this test which 90 percent of the examinees answer correctly would have an $RDR_i = 0.73$. If the item was revised so that $p = .80$, the new RDR_i would be 0.47. Items which are most extreme in difficulty should be revised first to improve $RDR_{\bar{X}}$ since the mean RDR_i for a test equals its $RDR_{\bar{X}}$.

The difficulty levels of content-parallel tests of equal length can be compared more efficiently with the $RDR_{\bar{X}}$. Data from Frisbie's (1973) study of multiple choice and true-false test reliabilities are reported in Table 1.

In his study multiple choice items were converted to true-false form by two different methods (judgmental and discrimination) in two subject matter areas (social studies and natural science). Subjects were randomly assigned to test forms. The $RDR_{\bar{X}}$ data in Table 1 show that in all eight comparisons, true-false

Insert Table 1 about here

subtest difficulty exceeded multiple choice subtest difficulty when each true-false subtest was paired with the multiple choice subtest from which it was derived. The relative difficulty of the true-false subtests probably is a major explanation for the relatively superior reliabilities found for the multiple choice subtests. Apparently when one converts a multiple choice item to true-false form, the latter is a somewhat more difficult item when chance success is adjusted.

The same author investigated the relative merits of the same two item forms by using lengthened true-false subtests to equate testing time (Frisbie, 1974). Random assignment of subjects to test forms was used to obtain group equivalence. The data in Table 2 show that subtest length was established to reflect a previous finding that three true-false items could be attempted in the time required to attempt a pair of multiple choice items. The values of $RDR_{\bar{X}}$ in the table show that true-false items were more difficult than multiple choice in each of the four comparisons. Obviously such comparisons should not be made without using $RDR_{\bar{X}}$ because of test length and chance score differences between tests.

Insert Table 2 about here

Item Difficulty

Item difficulties or p-values are generally-accepted indices used in test analysis and evaluation to assess the performance of specific test items.

Test specialists agree that items in norm-referenced tests should be of moderate difficulty: around 50 percent of the examinees should respond correctly. It is seldom pointed out, however, that the standard, 50 percent assumes a chance score of zero or the use of a correction for guessing formula. Since formula scoring is typically not used and not recommended (Ebel, 1974) for classroom testing, equation 7 should be used to define the standard for moderate difficulty for objective test items. In sum, the principle illustrated by equation 7 should be a part of general practice, not merely a part of common knowledge.

Items which were written to measure achievement of the same instructional objective and were written in different formats can be compared for differences in difficulty using RDR_i values obtained from the same group or comparable groups. This circumstance may evolve in conducting research on item formats or in selecting items for "equivalent" test forms for retesting purposes. Suppose that two content-parallel true-false and five-choice multiple choice items each have traditional p-values of .75. The respective RDR_i values for the two items are 0.00 and 0.38. The multiple choice item is clearly easier and, furthermore, the true-false item is at the difficulty level which will maximize item discrimination. Though validity, reliability, and testing time should be considered, on the basis of difficulty level alone, this true-false item is superior.

Differences in chance scores are frequently ignored when multiple choice items with varying number of response choices are evaluated. However, five-choice and three-choice items which have a traditional p-value of .80, for example, have RDR_i values of 0.50 and 0.39, respectively. Though this difference is relatively small, the five-choice item is in greater need of

attention than is the three-choice item. If items in a test are selected for revision or inclusion in a new test, the quality of the new instrument will be more greatly enhanced by using values of RDR_i than traditional p-values for selection purposes.

The sampling distribution of $\overline{RDR}_X$ warrants examination before strict guidelines can be given to describe extreme values. In addition, true differences between RDRs cannot be detected or tested statistically without investigating the properties of the sampling distribution. The RDR values given in Tables 1 and 2 may not be sufficiently discrepant to conclude that the true-false exceeded the multiple choice in difficulty. However, the data do support the conclusion that the multiple choice are not more difficult than the true-false.

References

Dryden, R.E. and Frisbie, D.A. Comparative reliabilities and validities of multiple choice and complex multiple choice nursing education tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Ebel, R.L. Essentials of education measurement (3rd ed.). Englewood Cliffs: Prentice-Hall, 1979.

Ebel, R.L. The ineffectiveness of multiple true-false items. Educational and Psychological Measurement, 1978, 38, 37-44.

Frisbie, D.A. Multiple choice versus true-false: A comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Frisbie, D.A. The effect of item format on reliability and validity: A study of multiple choice and true-false achievement tests. Educational and Psychological Measurement, 1974, 34, 885-892

Huck, S.W. Test performance under the condition of known item difficulty. Journal of Educational Measurement, 1978, 15, 53-58.

Hughes, H.H. and Trimble, W.E. The use of complex alternatives in multiple choice items. Educational and Psychological Measurement, 1965, 25, 117-126.

Table 1

Relative Test Difficulty for Content-Parallel MC and T-F Tests (Frisbie, 1973)

Form	N	$\bar{X}$	n	K	$RDR_{\bar{X}}$
SJA_m	126	24.31	4	35	0.18
SJB_t	127	24.48	2	35	-0.20
SJB_m	127	16.92	4	35	-0.38
SJA_t	126	19.55	2	35	-0.76
SDA_m	128	24.21	4	35	0.18
SDB_t	128	22.57	2	35	-0.42
SDB_m	128	17.44	4	35	-0.34
SDA_t	128	19.61	2	35	-0.76
NDA_m	125	22.19	4	35	0.02
NDB_t	129	19.36	2	35	-0.79
NDB_m	129	16.79	4	35	-0.39
NDA_t	125	21.00	2	35	-0.60
NJA_m	126	21.62	4	35	-0.02
NJB_t	129	20.78	2	35	-0.62
NJB_m	129	17.83	4	35	-0.31
NJA_t	126	19.71	2	35	-0.75

NOTE: Form codes are: S=social studies, N=natural science, J=judgmental conversion method, D=discrimination conversion method, and A and B refer to the order of the multiple choice (m) and true-false (t) subtests.

Table 2

Relative Test Difficulty for Content-Parallel MC and T-F Tests (Frisbie, 1974)

Form	N	$\bar{X}$	n	K	$RDR_{\bar{X}}$
SA_m	147	19.10	4	28	0.15
SB_t	148	26.92	2	42	-0.44
SB_m	148	12.80	4	28	-0.45
SA_t	147	24.09	2	42	-0.70
NA_m	117	17.57	4	28	0.01
NB_t	117	24.54	2	42	-0.66
NB_m	117	14.09	4	28	-0.32
NA_t	117	24.84	2	42	-0.63

NOTE: Form codes are: S=social studies, N=natural science, and A and B refer to the order of the multiple choice (m) and true-false (t) subtests.